**Note:** This document aims to be used as a guide to support the development of lrgext.x.py, to those intending to modify the script in a future and those looking for a reference of the steps and features to be accomplished during script development in a collaborative environment.

# PROGRAM DEVELOPMENT

### Initial framework and structure: IGP, Done

### Features:

1. Testing:
   1. Testing there is an allocated LRG for the given gene name: VF, Done
   2. Testing LRG file exists and is a readable file: IGP, Done
   3. Check that the xml version format (schema) is right: IGP, Done
   4. Check direction of strand and warn if in reverse: IGP, Done
   5. Check the number of builds and if more than one, check that coordinates are diff. VF, Done
   6. Check and compare the size of the reference build with the LRG sequence, VF, Done
   7. Data\_consistency: check that coordinates extracted are numbers: VF, IP
   8. Exon\_coor: check that exons do not overlap: IGP, IP
   9. Exon\_number: check that we get the right number of exons: IGP, Done
2. Usage:
   1. Extraction of input and outputs from different folders: IGP, Done
   2. Input gene name from command line: VF, Done
   3. Create a list of existing LRG files within directory: VF: Done
   4. Search LRG file for given gene: VF, Done
   5. Check and provide warnings: IGP + VF: Done
3. Parsing:
   1. Extract LRG background information: IGP, Done
   2. Extract gene name (later used to create LRG directory): VF, Done
   3. Extracting exon, transcript and protein coordinates per exon: IGP, Done
   4. Extract differences between builds: VF, Done
4. File production:
   1. Create .csv list of gene names and associated LRG IDs: VF, Done
   2. Create tab separated file: IGP, Done
   3. Create .csv file: IGP, Done
   4. Create bed file: IGP, Done
   5. Create .txt build mapper file: VF, Done
   6. Create .csv build differences file: VF, Done

### Disclaimer: IGP, Done

### Dealing with issues and bugs

1. Issue with LRG without protein coordinates in the first exon (e.g. LRG\_62 and LRG\_292): IGP, Done
2. Issue with some genes from list not being called: VF, IP
3. Issue with Linux-Windows incompatibilities: IGP, Done
4. Issue with gene list if not being created when not existing: IGP, Done
5. Issue with wrong number of exons when multiple transcripts: IGP, Done

* **Commenting**:

1. Initial commenting: IGP, Done
2. Update and revision: VF, Done

# Documentation:

### COPYRIGHT:

* 1. Policy checking (University and NHS): IGP, Done
  2. Structure and first drafting: IGP, Done
  3. Completion: IGP, Done
  4. Read proofing: VF, Done

### README:

1. Document structure: IGP, Done

2. Initial draft: IGP, Done

3. Content:

1. Global Background: VF, Done

2. Usage, installation: IGP, Done

4. Features description:

1. Main features: IGP, Done

2. Testing

- Background: VF, Done

- Test listing and description: IGP, Done

4. Parsing

- Background: VF, Done

- Features listing and description: IGP, Done

5. Reporting

- Background: VF, Done

- Features listing and Description: VF, Done

1. Versioning: IGP, Done
2. Disclaimer: IGP, done
3. Final editing and read-proofing: VF, pending completion of documentation

# OUTLOOK

1. Adding new test using 'assert': IGP, Done

2. Creating the name of the file automatically: IGP & VF, Done

3. Make software fully modular: IGP, Done

4. Access file from internet: IGP, deemed as not necessary for now, since LRG files will be downloaded/updated at periodic times to avoid delays due to internet connection

5. Incorporating get-ops: IGP, Outlook